arXiv:cond-mat/0609568v1 [cond-mat.dis-nn] 22 Sep 2006

Statistical Mechanics of Linear and Nonlinear Time-Domain Ensemble Learning

Seiji MIYOSHI[1] * and Masato OKADA[234]

[1] *Department of Electronic Engineering, Kobe City College of Technology, 8–3 Gakuenhigashimachi, Nishi-ku, Kobe-shi, 651–2194*

[2] *Division of Transdisciplinary Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5–1–5 Kashiwanoha, Kashiwa-shi, Chiba, 277–8561*

[3] *RIKEN Brain Science Institute, 2–1 Hirosawa, Wako-shi, Saitama, 351–0198*

Conventional ensemble learning combines students in the space domain. In this paper, however, we combine students in the time domain and call it time-domain ensemble learning. We analyze, compare, and discuss the generalization performances regarding time-domain ensemble learning of both a linear model and a nonlinear model. Analyzing in the framework of online learning using a statistical mechanical method, we show the qualitatively different behaviors between the two models. In a linear model, the dynamical behaviors of the generalization error are monotonic. We analytically show that time-domain ensemble learning is twice as effective as conventional ensemble learning. Furthermore, the generalization error of a nonlinear model features nonmonotonic dynamical behaviors when the learning rate is small. We numerically show that the generalization performance can be improved remarkably by using this phenomenon and the divergence of students in the time domain.

KEYWORDS: ensemble learning, online learning, generalization error, statistical mechanics

1. Introduction

Learning can be roughly classified into batch learning and online learning.[1] In batch learning, given examples are used more than once. In this paradigm, a student will give the correct answers after training if that student has an adequate degree of freedom. However, it is necessary to have plenty of time and a large memory for storing many examples. On the contrary, in online learning examples used once are then discarded. In this case, a student cannot give correct answers for all the examples used in training. There are, however, merits. For example, a large memory for storing many examples is not necessary, and it is possible to follow a time-variant teacher.

Recently, we analyzed the generalization performance of some models in a framework of online learning using a statistical mechanical method.[2–6] Ensemble learning means to combine many rules or learning machines (called students in this paper) that perform poorly; this has recently attracted the attention of many researchers.[2,3,7–10] The diversity or variety of

*E-mail address: miyoshi@kobe-kosen.ac.jp

students is essential in ensemble learning. We showed that the three well-known rules, Hebbian learning, perceptron learning, and AdaTron learning have different characteristics in their affinities for ensemble learning, that is in "maintaining diversity among students"[3]D In that process,[13,14] it was subsidiarily proven that in an unlearnable case,[11,12] the student vector does not converge in one direction but continues moving. Therefore, we also analyzed the generalization performance of a student supervised by a moving teacher that goes around a true teacher,[4] proving that the generalization error of a student can be smaller than a moving teacher's, even if the student only uses examples from the moving teacher. In an actual human society, a teacher observed by a student does not always present the correct answer. In fact, many cases the teacher is learning and continues to change. Therefore, analyzing such a model is interesting in terms of considering the analogies between statistical learning theories and real human society.

In conventional ensemble learning, generalization performance is improved by combining students who have diversities. However, students do not always converge in one direction but may continue moving in an unlearnable model. Therefore, generalization performance in such a model must be improved by combining students themselves at different times, even if there is only one student. Conventional ensemble learning combines students in the space domain. In contrast, we introduce a method of combining the students in the time domain, which we call "time-domain ensemble learning".[6]

Some studies[15–18] have treated the combining of students in the time domain. We particularly pay attention to dynamical behaviors of the generalization performance of the time-domain ensemble learning and theoretically analyze it by applying a statistical mechanical method. We analytically or numerically obtain, compare, and discuss the order parameters and the generalization errors of two models: a linear model in which both teacher and student are linear perceptrons[2] with noise and a nonlinear model in which both teacher and student are nonlinear perceptrons. The results show that the two models have the qualitatively different behaviors. We analytically demonstrate that time-domain ensemble learning of a linear model is twice as effective as conventional ensemble learning. Furthermore, we numerically show that the generalization performance of a nonlinear model can be ramarkably improved by using nonmonotonic dynamical behaviors and the divergence of students in the time domain.

2. Model

In this paper we consider a teacher and a student. They are perceptrons with the connection weights $\boldsymbol{B}$ and $\boldsymbol{J}^m$, respectively, where m denotes the time step. For simplicity, the connection weights of the teacher and the student are simply called the teacher and the student. Teacher $\boldsymbol{B} = (B_1, \ldots, B_N)$, student $\boldsymbol{J}^m = (J_1^m, \ldots, J_N^m)$, and input $\boldsymbol{x}^m = (x_1^m, \ldots, x_N^m)$ are N-dimensional vectors. Each component B_i of $\boldsymbol{B}$ is independently drawn from $\mathcal{N}(0,1)$ and fixed, where $\mathcal{N}(0,1)$ denotes a Gaussian distribution with a mean of zero and a variance of

unity. Each component J_i^0 of the initial value $\boldsymbol{J}^0$ of $\boldsymbol{J}^m$ is independently drawn from $\mathcal{N}(0,1)$. The direction cosine between $\boldsymbol{J}^m$ and $\boldsymbol{B}$ is R^m and that between $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$ is $q^{m,m'}$. Each component x_i^m of $\boldsymbol{x}^m$ is drawn from $\mathcal{N}(0,1/N)$ independently.

Figure 1 illustrates the relationship among teacher $\boldsymbol{B}$, students $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$ and the direction cosines $R^m, R^{m'}$, and $q^{m,m'}$.

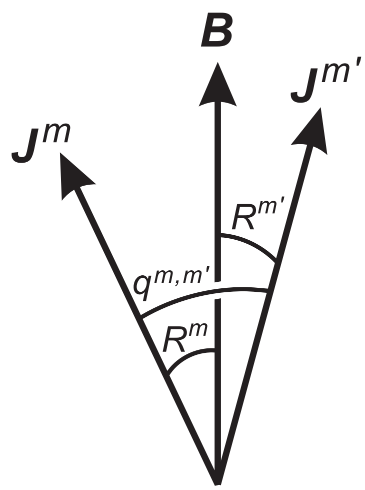

Fig. 1. Teacher $\boldsymbol{B}$ and students $\boldsymbol{J}^m$ and $\boldsymbol{J}^{m'}$. $R^m, R^{m'}$, and $q^{m,m'}$ are direction cosines.

In this paper, we also deal with the thermodynamic limit $N \to \infty$. Therefore, $\|\boldsymbol{B}\| = \sqrt{N}$, $\|\boldsymbol{J}^0\| = \sqrt{N}$, and $\|\boldsymbol{x}^m\| = 1$. Generally, since the norm $\|\boldsymbol{J}^m\|$ of the student changes as the time step proceeds, the ratios l^m of the norm to $\sqrt{N}$ are introduced and are called the length of the student. That is, $\|\boldsymbol{J}^m\| = l^m\sqrt{N}$.

Linear Case Outputs of the teacher and the student are $o_B^m = v^m + n_B^m$ and $o_J^m = u^m l^m + n_J^m$, respectively. Here, $v^m = \boldsymbol{B}\cdot\boldsymbol{x}^m, u^m l^m = \boldsymbol{J}^m\cdot\boldsymbol{x}^m, n_B^m \sim \mathcal{N}\left(0,\sigma_B^2\right), n_J^m \sim \mathcal{N}\left(0,\sigma_J^2\right)$, where $\mathcal{N}(0,\sigma^2)$ denotes a Gaussian distribution with a mean of zero and a variance of σ^2. That is, the outputs of the teacher and the student include independent Gaussian noises with variances of σ_B^2 and σ_J^2, respectively. Then, v^m and u^m obey Gaussian distributions with a mean of zero, a variance of unity, and a covariance of R^m. Let us define the error ϵ_S^m between the teacher $\boldsymbol{B}$ and the student $\boldsymbol{J}^m$ alone by the squared error of their outputs:

$$\epsilon_S^m \equiv \frac{1}{2}\left(o_B^m - o_J^m\right)^2. \tag{1}$$

Student $\boldsymbol{J}^m$ adopts the gradient method as a learning rule and uses input $\boldsymbol{x}$ and an output of teacher $\boldsymbol{B}$ for updates. That is,

$$\begin{aligned} \boldsymbol{J}^{m+1} &= \boldsymbol{J}^m - \eta\frac{\partial\epsilon_S^m}{\partial\boldsymbol{J}^m} && (2)\\ &= \boldsymbol{J}^m + \eta\left(v^m + n_B^m - u^m l^m - n_J^m\right)\boldsymbol{x}^m, && (3) \end{aligned}$$

where η denotes the learning rate of the student and is a constant positive number. The part $\eta\left(v^m + n_B^m - u^m l^m - n_J^m\right)$ has been determined by the learning rule. Generalizing this part, we denote it with f^m.

Nonlinear Case The teacher and the student are a nonmonotonic perceptron and a simple perceptron, respectively; their outputs are $o_B^m = \mathrm{sgn}\left(\left(v^m - a\right) v^m \left(v^m + a\right)\right),\ o_J^m = \mathrm{sgn}\left(u^m l^m\right)$. Here, $v^m = \boldsymbol{B}\cdot\boldsymbol{x}^m$ and $u^m l^m = \boldsymbol{J}^m \cdot \boldsymbol{x}^m$, v^m and u^m obey Gaussian distributions with a mean of zero, a variance of unity, and a covariance of R^m, and $\mathrm{sgn}(\cdot)$ denotes a sign function. Student $\boldsymbol{J}^m$ adopts the perceptron learning as a learning rule for updates. That is,

$$\boldsymbol{J}^{m+1} = \boldsymbol{J}^m + \eta\Theta\left(-o_B^m o_J^m\right) o_B^m \boldsymbol{x}^m, \tag{4}$$

where η denotes the learning rate of the student and is a constant positive number, $\Theta(\cdot)$ denotes a step function. The part $\eta\Theta\left(-o_B^m o_J^m\right) o_B^m$ has been determined by the learning rule. Generalizing this part, we denote it with f^m.

3. Theory

3.1 Generalization error

Conventionally, ensemble learning means to improve performance by combining many students that perform poorly. We, however, use just one student and combine copies of it (hereafter called "brothers") at different time steps in this paper. Conventional ensemble learning combines students in the space domain, whereas, we do so in the time domain. In this paper K brothers $\boldsymbol{J}^{m_1}, \boldsymbol{J}^{m_2}, \ldots, \boldsymbol{J}^{m_K}$ are combined. Here, $m_1 \le m_2 \le \ldots \le m_K$. One goal of statistical learning theory is to theoretically obtain generalization errors. Since a generalization error is the mean of errors over the distribution of the new input $\boldsymbol{x}$, the generalization errors ϵ_g of the ensemble in linear and nonlinear cases are calculated as follows:

Linear Case We use the squared error ϵ for new input $\boldsymbol{x}$. Here, it is assumed that the Gaussian noises are independently added to the teacher and each brother of the ensemble. The weight of each brother $\boldsymbol{J}^{m_k}$ of the ensemble satisfies $C_k \ge 0$. That is, the error of the ensemble is

$$\epsilon = \frac{1}{2}\left(\boldsymbol{B}\cdot\boldsymbol{x} + n_B - \sum_{k=1}^{K} C_k\left(\boldsymbol{J}^{m_k}\cdot\boldsymbol{x} + n_k\right)\right)^2, \tag{5}$$

where $n_B \sim \mathcal{N}\left(0, \sigma_B^2\right)$ and $n_k \sim \mathcal{N}\left(0, \sigma_J^2\right)$. Thus, the generalization error ϵ_g of the ensemble is calculated as follows:

$$\begin{aligned}\epsilon_g &= \int d\boldsymbol{x} dn_B \left(\prod_{k=1}^{K} dn_k\right) p(\boldsymbol{x}) p(n_B) \left(\prod_{k=1}^{K} p(n_k)\right)\epsilon \qquad (6)\\ &= \int dv \left(\prod_{k=1}^{K} du_k\right) dn_B \left(\prod_{k=1}^{K} dn_k\right) p(v, \{u_k\}) p(n_B)\left(\prod_{k=1}^{K} p(n_k)\right)\end{aligned}$$

$$\times\frac{1}{2}\left(v+n_B-\sum_{k=1}^{K}C_k\left(u_k l^{m_k}+n_k\right)\right)^2 \tag{7}$$

$$= \frac{1}{2}\left(1-2\sum_{k=1}^{K}C_k l^{m_k}R^{m_k}+2\sum_{k=1}^{K}\sum_{k'>k}^{K}C_kC_{k'}l^{m_k}l^{m_{k'}}q^{m_k,m_{k'}}\right.$$
$$\left.+\sum_{k=1}^{K}C_k^2(l^{m_k})^2+\sigma_B^2+\sum_{k=1}^{K}C_k^2\sigma_J^2\right), \tag{8}$$

where $v=\boldsymbol{B}\cdot\boldsymbol{x}$ and $u_k l^{m_k}=\boldsymbol{J}^{m_k}\cdot\boldsymbol{x}$. We executed integration using the following: v and u_k obey $\mathcal{N}(0,1)$. The covariance between v and u_k is R^{m_k}, and that between u_k and $u_{k'}$ is $q^{m_k,m_{k'}}$.

Nonlinear Case A majority vote by brothers might be a typical method of combining for a nonlinear model in which the output of each student is $+1$ or -1. However, to simplify the analysis we apply the following method: the output of the ensemble is that of a new perceptron of which the connection weight is the weighted sum of the normalized connection weights $\boldsymbol{J}^{t_k}/l^{t_k}$ of brothers. That is,

$$\epsilon=\Theta\left(-(v-a)\,v\,(v+a)\sum_{k=1}^{K}C_k u_k\right), \tag{9}$$

where $C_k\geq 0$ is a weight of each brother $\boldsymbol{J}^{m_k}$ in the ensemble. Thus, the generalization error ϵ_g of the ensemble is expressed as follows:

$$\epsilon_g=\int d\boldsymbol{x}p(\boldsymbol{x})\epsilon=\int dv\left(\prod_{k=1}^{K}du_k\right)p(v,\{u_k\})\Theta\left(-(v-a)\,v\,(v+a)\sum_{k=1}^{K}C_k u_k\right). \tag{10}$$

3.2 Differential equations for order parameters, and their solutions

In this paper, we examine the thermodynamic limit $N\to\infty$. To do so, updates of Eq.(3) or Eq.(4) must be executed $O(N)$ times for the order parameters l, R, and q to change by $O(1)$. Thus, the continuous times $t_1,\ldots,t_K$, which are the time steps $m_1,\ldots,m_K$ normalized by the dimension N, are introduced as the superscripts that represent the learning process. To simplify the analysis, we introduced the following auxiliary order parameters $r^t\equiv l^tR^t$ and $Q^{t,t'}\equiv l^tl^{t'}q^{t,t'}$. The simultaneous differential equations in deterministic forms,[19] which describe the dynamical behaviors of order parameters, have been obtained based on the self-averaging of thermodynamic limits as follows:

$$\frac{dl^t}{dt} = \langle f^tu^t\rangle+\frac{\langle (f^t)^2\rangle}{2l^t}, \tag{11}$$

$$\frac{dr^t}{dt} = \langle f^tv^t\rangle, \tag{12}$$

$$\frac{dQ^{t,t'}}{dt'} = l^t\langle f^{t'}\bar{u}^t\rangle, \tag{13}$$

where $t' \geq t$ and $\bar{u}^t = \boldsymbol{x}^{t'} \cdot \boldsymbol{J}^t / l^t \sim \mathcal{N}(0,1)$. Four sample averages in Eqs.(11)–(13) are obtained by executing integrations where $v^{t'}, u^{t'}$ and $\bar{u}^t$ obey the triple-Gaussian distribution $p(v^{t'}, u^{t'}, \bar{u}^t)$, for which the covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & R^{t'} & R^t \\ R^{t'} & 1 & q^{t,t'} \\ R^t & q^{t,t'} & 1 \end{pmatrix}. \tag{14}$$

Linear Case The four sample averages can be easily calculated as follows:

$$\langle f^t u^t \rangle = \eta(r^t/l^t - l^t), \tag{15}$$

$$\langle (f^t)^2 \rangle = \eta^2(1 + \sigma_B^2 + \sigma_J^2 + (l^t)^2 - 2r^t), \tag{16}$$

$$\langle f^t v^t \rangle = \eta(1 - r^t), \tag{17}$$

$$\langle f^{t'} \bar{u}^t \rangle = \eta \left(r^t - Q^{t,t'} \right) / l^t. \tag{18}$$

Using $R^0 = 0, l^0 = 1$ and $Q^{t,t} = (l^t)^2$ as initial conditions, we can analytically solve the simultaneous differential equations Eqs.(11)–(13) as follows:[6]

$$r^t = 1 - e^{-\eta t}, \tag{19}$$

$$(l^t)^2 = 1 + \frac{\eta}{2-\eta}\left(\sigma_B^2 + \sigma_J^2\right) - 2e^{-\eta t} + \left(2 - \frac{\eta}{2-\eta}\left(\sigma_B^2 + \sigma_J^2\right)\right) e^{\eta(\eta-2)t}, \tag{20}$$

$$Q^{t,t'} = 1 - e^{-\eta t} + e^{-\eta t'} + \left((l^t)^2 - 1\right) e^{-\eta(t'-t)}. \tag{21}$$

Substituting Eqs.(19)–(21) into Eq.(8), the generalization error ϵ_g can be analytically obtained as a function of time $t_k,\ k = 1, \ldots, K$.

Nonlinear Case The four sample averages are obtained as follows:

$$\langle f^t u^t \rangle = \frac{\eta}{\sqrt{2\pi}} \left(R^t \left(2\exp\left(-\frac{a^2}{2}\right) - 1 \right) - 1 \right), \tag{22}$$

$$\langle (f^t)^2 \rangle = 2\eta^2 \left(\int_a^\infty Dv H\left(\frac{R^t v}{\sqrt{1-(R^t)^2}} \right) + \int_0^a Dv H\left(-\frac{R^t v}{\sqrt{1-(R^t)^2}} \right) \right), \tag{23}$$

$$\langle f^t v^t \rangle = \frac{\eta}{\sqrt{2\pi}} \left(2\exp\left(-\frac{a^2}{2}\right) - 1 - R^t \right), \tag{24}$$

$$\langle f^{t'} \bar{u}^t \rangle = \frac{1}{1-(R^{t'})^2} \left(\left(q^{t,t'} - R^{t'} R^t \right) \langle f^{t'} u^{t'} \rangle + \left(R^t - q^{t,t'} R^{t'} \right) \langle f^{t'} v^{t'} \rangle \right), \tag{25}$$

where $H(u) \equiv \int_u^\infty Dx,\ Dx \equiv \frac{dx}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. The generalization error ϵ_g is obtained by solving Eqs.(10)–(13), and (22)–(25) numerically.

4. Results and Discussion

Figure 1 shows examples of the dynamical behaviors of l and R in a linear model obtained analytically, and the corresponding simulation results, where $N = 2,000$. In the case of a

linear model, many properties regarding both dynamical behaviors and steady states can be analytically proven.[6] For example, l and ϵ_g diverge unless $0 < \eta < 2$. In the case of no noise, l asymptotically approaches unity after becoming larger than unity if $0 < \eta < 1$ and l asymptotically approaches unity after becoming smaller than unity if $1 < \eta < 2$. The larger η is, the faster R rises. However, the convergence of R is fastest when the learning rate satisfies $\eta = 1$. This phenomenon can be understood by the fact that $\eta = 1$ is a special condition for the gradient method where the student uses up the information obtained from input $\boldsymbol{x}$.[5]

Figure 2 displays some examples of the dynamical behaviors of l and R in a nonlinear model obtained numerically, and the corresponding simulation results, where $N = 2,000$. The reason why R is negative, which differs from the linear case, is that the threshold a of a nonmonotonic teacher is greater than the critical value $a_C = \sqrt{2\ln 2} \simeq 1.18$.[11] This is not essential in this paper. When the learning rate η is relatively large, the dynamical behavior of R is monotonic; however, when η is small, the dynamical behavior of R becomes nonmonotonic. That is, $|R|$ asymptotically approaches a steady value after reaching its maximal one. The steady value is not dependent upon η.

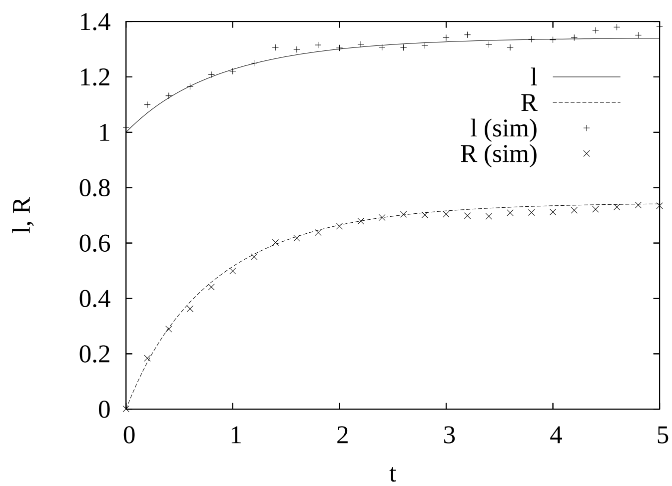

Fig. 2. Dynamical behaviors of l and R in the linear case. $\eta = 1.0,\ \sigma_B^2 = 0.3,\ \sigma_J^2 = 0.5,$

Figure 4 (left) presents some examples of the dynamical behaviors of q of a linear model obtained analytically, and the corresponding simulation results, where $N = 2,000$. For a linear model, q increases monotonically when t increases, and increases monotonically when $t' - t$ decreases. Furthermore, $q^{t,t'}$ converges to a smaller value than unity in the case of $t' - t \neq 0.0$. This means the student itself continues to move after the order parameters reach a steady state. Figure 4 (right) shows the relationship between t_1 and ϵ_g in the case of constant $t_2 - t_1$ and $K = 2$. Here, ϵ_g increases monotonically, remains constant, or decreases monotonically depending on the values of η.

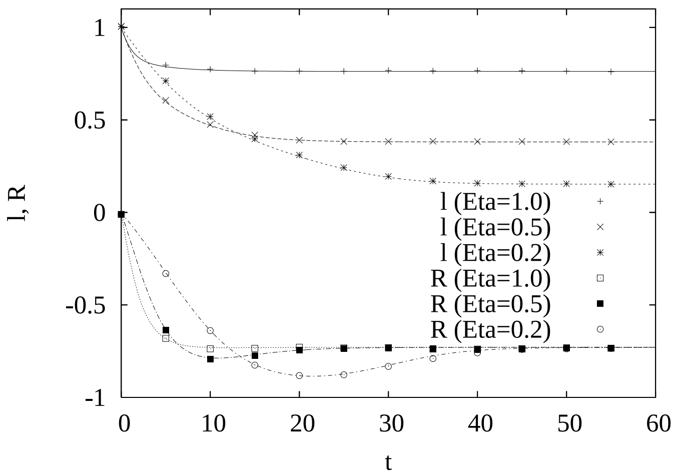

Fig. 3. Dynamical behaviors of l and R in the nonlinear case. $a = 2.0$.

The behaviors of ϵ_g when the leading time $t_1 \to \infty$ and the time interval $t_{k+1} - t_k \to \infty$ can be theoretically obtained in the case of a linear model as follows:[6] ϵ_g decreases as η decreases regardless of K. When the weights are uniform or $C_k = C = 1/K$ and $K = \infty$, $\boldsymbol{B} = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{J}^{t_k}$. This means the generalization error equals the residual error caused by teacher's noise n_B. On the other hand, the generalization error ϵ_g of $K = \infty$ is $\frac{1}{4}$ times of that of $K = 1$ when the learning rate satisfies $\eta = 1$, the uniform weights $C_k = 1/K$, $\sigma_B^2 = \sigma_J^2$, $t_1 \to \infty$, and $t_{k'} - t_k \to \infty$. Since the generalizaion error ϵ_g of the conventional space-domain ensemble learning with $K = \infty$, $\eta = 1$, $C_k = 1/K$ and $\sigma_B^2 = \sigma_J^2$ is $\frac{1}{2}$ times of that of $K = 1$,[2] we can say the time-domain ensemble learning is twice as effective as the conventional space-domain ensemble learning. This difference can be explained as follows: In conventional space-domain ensemble learning, the similarities among students become high since all students use the same examples for learning. In time-domain ensemble learning, on the other hand, the similarities among brothers become low since all brothers use almost totally different examples for learning.

Figures 5–7 (left) show some examples of the dynamical behaviors of q for a nonlinear model obtained numerically, and the corresponding simulation results, where $N = 2,000$. These figures indicate that q for a nonlinear model behaves nonmonotonically for t when η is small. Figures 5–7 (right) show the relationship between t_1 and ϵ_g in the case of constant $t_2 - t_1$ and $K = 2$. The steady value of ϵ_g is dependent upon $t_2 - t_1$ and is not dependent upon η. However, when η is small, ϵ_g behaves nonmonotonically for t_1 and has the minimal value shown in Figs.6 (right) and 7 (right). This phenomenon can be considered a kind of over-learning. Figure 7 (right) shows that the minimal value of ϵ_g decreases when $t_2 - t_1$ increases as $0 \to 5 \to 10$ and increases when $t_2 - t_1$ increases as $10 \to 20$. This means that $t_2 - t_1$ has an optimum value. Figures 6 (right) and 7 (right) reveal that the smaller the learning rate η is,

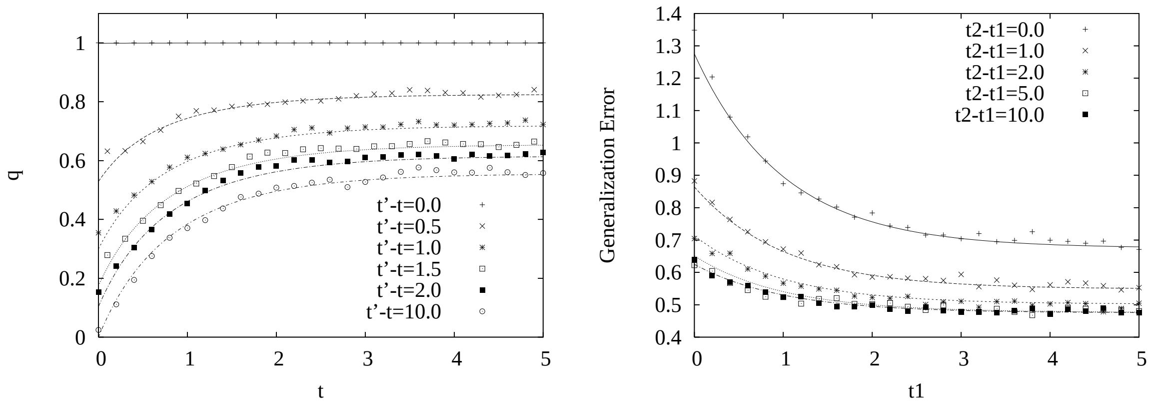

Fig. 4. Dynamical behaviors of q and ϵ_g in a linear case. $\eta = 1.0,\ \sigma_B^2 = 0.3,\ \sigma_J^2 = 0.5,\ K = 2,\ C_k = 1/K$.

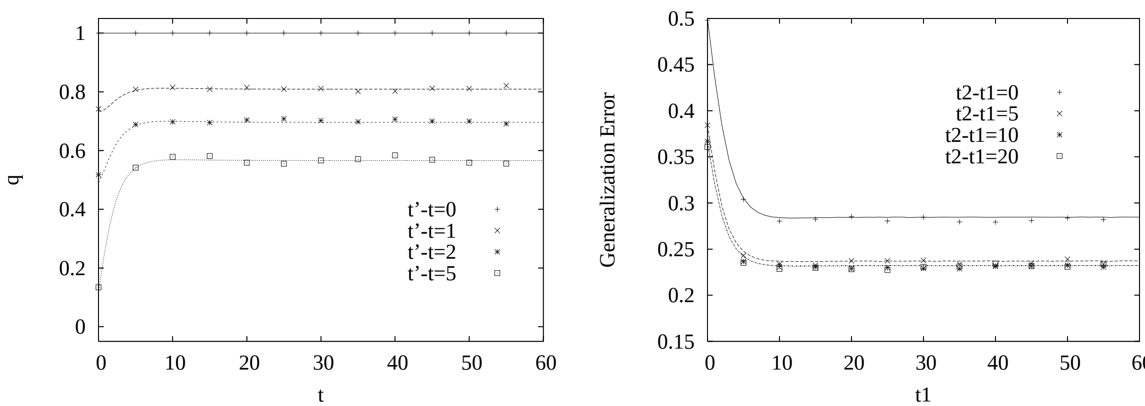

Fig. 5. Dynamical behaviors of q and ϵ_g in a nonlinear case. $\eta = 1.0$, $a = 2.0,\ K = 2,\ C_k = 1/K$.

the smaller the minimal value of ϵ_g is. However, if η is too small, the learning becomes slow. Therefore, if the aim is to decrease the generalization error ϵ_g, we should use the smallest η that is possible from the viewpoint of learning speed, set $t' - t$ to the optimum value, and stop the learning at an adequate time step.

5. Conclusion

We have analyzed the generalization performances regarding time-domain ensemble learning of both a linear model and a nonlinear model. Analyzing within the framework of online learning using a statistical mechanical method, we have demonstrated the qualitatively different behaviors between the two models. In a linear model, the dynamical behaviors of the generalization error are monotonic. We have analytically shown that time-domain ensemble learning is twice as effective as conventional ensemble learning. Furthermore, the generalization error of a nonlinear model exhibits nonmonotonic dynamical behaviors when the learning rate is small. We have numerically shown that the generalization performance can be remark-

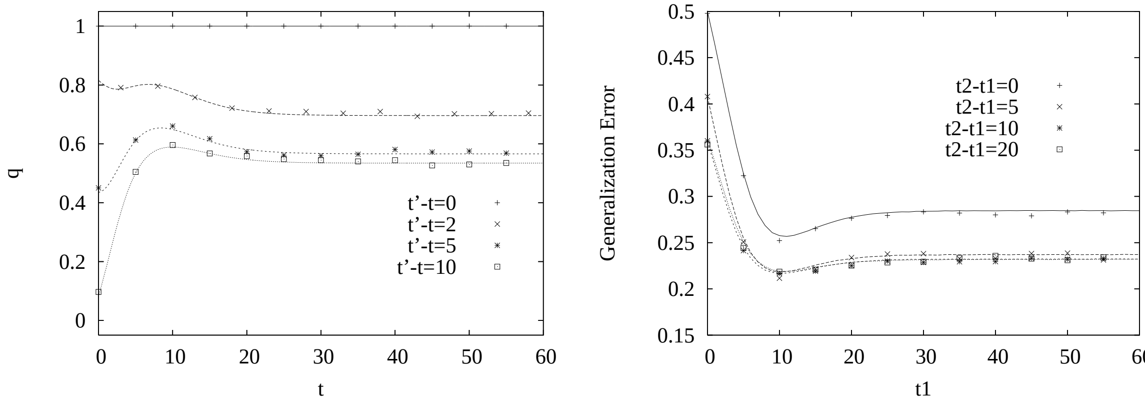

Fig. 6. Dynamical behaviors of q and ϵ_g in a nonlinear case. $\eta = 0.5$, $a = 2.0$, $K = 2$, $C_k = 1/K$.

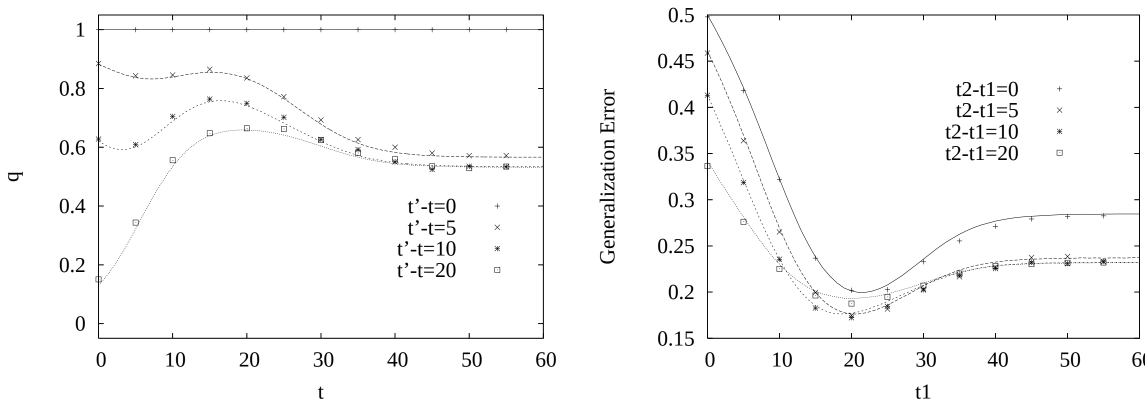

Fig. 7. Dynamical behaviors of q and ϵ_g in a nonlinear case. $\eta = 0.2$, $a = 2.0$, $K = 2$, $C_k = 1/K$.

ably improved by using this phenomenon together with the divergence of students in the time domain.

Acknowledgments

This research was partially supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan, with Grants-in-Aid for Scientific Research 14084212, 15500151, 16500093, and 18500183.

References

1) Saad, D. (eds.) (1998) *On-line Learning in Neural Networks.* Cambridge: Cambridge University Press.
2) Hara, K. & Okada, M. (2005) Ensemble learning of linear perceptron; Online learning theory. *Journal of Physical Society of Japan* **74**:2966-2972.
3) Miyoshi, S., Hara, K. & Okada, M. (2005) Analysis of ensemble learning using simple perceptrons based on online learning theory. *Physical Review E* **71**:036116.
4) Miyoshi, S. & Okada, M. (2006) Analysis of on-line learning when a moving teacher goes around a true teacher. *Journal of Physical Society of Japan* **75**(2):024003.
5) Miyoshi, S. & Okada, M. (2006) Statistical mechanics of online learning for ensemble teachers. *Journal of Physical Society of Japan* **75**(4):044002.
6) Miyoshi, S., Uezu, T. & Okada, M. (2006) Statistical mechanics of time domain ensemble learning. *Journal of Physical Society of Japan* **75**(8):084007.
7) Freund, Y. & Schapire, R.E. (1999) A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* **14**:771-780. [in Japanese, translation by Abe, N.]
8) http://www.boosting.org/
9) Statistical mechanics of ensemble learning. Krogh, A. & Sollich, P. (1997) *Physical Review E* **55**:811-825.
10) Online learning with ensembles. Urbanczik, R. (2000) *Physical Review E* **62**:1448-1451.
11) Inoue, J.I. & Nishimori, H. (1997) On-line AdaTron learning of unlearnable rules. *Physical Review E* **55**:4544-4551.
12) Inoue, J.I., Nishimori, H. & Kabashima, Y. (1997) A simple perceptron that learns non-monotonic rules. cond-mat/9708096.
13) Miyoshi, S., Hara, K. & Okada, M. (2004) Analysis of ensemble learning for committee machine teacher. *Proc. The Seventh Workshop on Information-Based Induction Sciences*: 178-185. [in Japanese]
14) Miyoshi, S., Hara, K. & Okada, M. (2005) Analysis of ensemble learning for non-monotonic teacher. *IEICE Technical Report* **NC2004-214**:123-128. [in Japanese]
15) Dekel, O. & Singer, Y. (2005) Data-driven online to batch conversions. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 18.*
16) Freund, Y. & Shapire, R.E. (1999) Large margin classification using the perceptron algorithm. *Machine Learning* **37**(3):277-296.
17) Cesa-Bianchi, N., Conconi, A. & Gnetile, C. (2004) On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory* **50**(9):2050-2057.
18) Cesa-Bianchi, N. & Gnetile, C. (2005) Improved risk tail bounds for on-line algorithms. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 18.*
19) Nishimori, H. (2001) *Statistical Physics of Spin Glasses and Information Processing: An Introduction.* Oxford: Oxford University Press.